Language Technology

Hartmut Fitz

Department of Information Science
University of Groningen
Fall 2009/10
h.fitz@rug.nl

September 21, 2009

Overview

Part II

- Basic syntactic development
- Learning word categories
- Learning to order words
- Computational modelling
- Psycholinguistics ∞ language technology

Word categories

Assigning word categories critical for comprehension

Temporary ambiguity: The boys who eat fish
N

Word categories

Assigning word categories critical for comprehension

Temporary ambiguity: The boys who eat fish by the lake.
V

Word categories

Assigning word categories critical for comprehension

Temporary ambiguity: The boys who eat fish by the lake.
V

Global ambiguity: Flying planes made her duck.

Word categories

Assigning word categories critical for comprehension

Temporary ambiguity: The boys who eat fish by the lake.
V

Global ambiguity: Flying planes made her duck.
V N V

Word categories

Assigning word categories critical for comprehension

Temporary ambiguity: The boys who eat fish by the lake.
V

Global ambiguity: Flying planes made her duck.
Adj N N

Word categories

Assigning word categories critical for comprehension

Temporary ambiguity: The boys who eat fish by the lake.
V

Global ambiguity: Flying planes made her duck.
Adj N V

Word categories

Assigning word categories critical for comprehension

Temporary ambiguity: The boys who eat fish by the lake.
V

Global ambiguity: Flying planes made her duck.
Adj N V

Examples illustrate

- Words can be in multiple classes
- Incremental processing, online assignment
- Non-monotonic: computation and re-computation of meaning

What are word categories?

"To understand how X is learned, you first have to understand what X is." (Pinker, 1990)

Major word categories

Nouns	objects, things
Verbs	processes, actions, states
Adjectives	properties of object
Prepositions	relations between objects (e.g., spatial)
Adverbs	modify verbs
Pronoun	substitutes for nouns, marked for person
⋮	⋮

What are word categories?

Yes, but...

What are word categories?

Yes, but...

Fuzzy criteria	Abstracta (belief), events (earthquake) states (depression), qualities (strength)

What are word categories?

Yes, but...

Fuzzy criteria	Abstracta (belief), events (earthquake) states (depression), qualities (strength)
Context-dependence	He's staggering/his staggering/ his staggering wealth

What are word categories?

Yes, but...

Fuzzy criteria	Abstracta (belief), events (earthquake) states (depression), qualities (strength)
Context-dependence	He's staggering/his staggering/ his staggering wealth
Theory-dependence	No two syntactic theories agree on taxonomy of word classes

What are word categories?

Yes, but...

Fuzzy criteria	Abstracta (belief), events (earthquake) states (depression), qualities (strength)
Context-dependence	He's staggering/his staggering/ his staggering wealth
Theory-dependence	No two syntactic theories agree on taxonomy of word classes
Language-dependence	Stative verbs and adjectives difficult to distinguish in Chinese
	Two classes of adjectives in Japanese
	No one-one mapping between languages

What are word categories?

Distributional properties

- Structuralism: conceptual (semantic) definitions are vacuous (Palmer, 1971)
- Word categories should be defined by distributional properties
- Words assigned to class based on occurrence in similar syntactic frames (e.g., X is VERB-ing Y)
- Today: word categories based on various cues, including
  - phonological and morphological properties of words
  - distributional information
  - semantic features

What are word categories?

Distributional properties

- Structuralism: conceptual (semantic) definitions are vacuous (Palmer, 1971)
- Word categories should be defined by distributional properties
- Words assigned to class based on occurrence in similar syntactic frames (e.g., X is VERB-ing Y)
- Today: word categories based on various cues, including
  - phonological and morphological properties of words
  - distributional information
  - semantic features

Distributional information

Some psycholinguistic evidence

Distributional information

Some psycholinguistic evidence

- Artificial grammar learning: children can learn non-adjacent dependencies in syntactic frame-like word chunks (Gomez, 2002)

Distributional information

Some psycholinguistic evidence

- Artificial grammar learning: children can learn non-adjacent dependencies in syntactic frame-like word chunks (Gomez, 2002)
- Children can abstract word categories from distributional cues in speech (Gerken, Wilson & Lewis, 2005)

Distributional information

Some psycholinguistic evidence

- Artificial grammar learning: children can learn non-adjacent dependencies in syntactic frame-like word chunks (Gomez, 2002)
- Children can abstract word categories from distributional cues in speech (Gerken, Wilson & Lewis, 2005)
- Children acquire novel verbs more easily when they occur in syntactic frames that are frequent in the input (Childers & Tomasello, 2001)

Distributional information

Research questions beyond AGL

- What type of distributional information in natural speech is particularly informative?
- What kinds of distributional cues are infants sensitive to in categorizing words?
- How can distributionally defined categories be integrated into grammatical system?
- Which concrete mechanisms of statistical learning are used? (Building computational models especially useful here)

Frequent frames (Mintz 2003 & 2006)

Basic idea

- Data: corpora of child-directed speech (individual children)
- Define frame as ordered triple X W Y: word W in context X Y
- If frame occurs frequently in corpus, this might be caused by some systematic aspect of language
- Likely to reflect some relationship between the W in frame, e.g., joint word category membership
- Measure/examine how predictive frames are for category membership

Potential problems

Multiple categories

Potential problems

Multiple categories

(a) Tom ate fish.
(b) Tom ate rabbits.

Syntactic frames: categorize fish and rabbit together

Potential problems

Multiple categories

(a) Tom ate fish.
(b) Tom ate rabbits.
(c) Tom can fish.
(d) *Tom can rabbits.

Syntactic frames: categorize fish and rabbit together

Leads to incorrect generalization

Potential problems

Multiple categories

(a) Tom ate fish.
(b) Tom ate rabbits.
(c) Tom can fish.
(d) *Tom can rabbits.

Syntactic frames: categorize fish and rabbit together

Leads to incorrect generalization

Non-local information

Potential problems

Multiple categories

(a) Tom ate fish.
(b) Tom ate rabbits.
(c) Tom can fish.
(d) *Tom can rabbits.

Syntactic frames: categorize fish and rabbit together

Leads to incorrect generalization

Non-local information

(a) to X to

X likely from same category verb

Potential problems

Multiple categories

(a) Tom ate fish.
(b) Tom ate rabbits.
(c) Tom can fish.
(d) *Tom can rabbits.

Syntactic frames: categorize fish and rabbit together

Leads to incorrect generalization

Non-local information

(a) to X to
(b) to quickly X to

X likely from same category verb

Split infinitive disrupts frame

Potential problems

Multiple categories

(a) Tom ate fish.
(b) Tom ate rabbits.
(c) Tom can fish.
(d) *Tom can rabbits.

Syntactic frames: categorize fish and rabbit together

Leads to incorrect generalization

Non-local information

(a) to X to
(b) to quickly X to

X likely from same category verb
Split infinitive disrupts frame

Do these issues undermine usefulness of distributional information?

Frequent frames

Procedure

- 6 corpora selected from CHILDES
- All frames X W Y are counted (separately by corpus)
- 45 most frequent frames selected (from one corpus)
  - you __ it | the __ and | put __ in | ...
- W from each occurrence of X W Y in each corpus are recorded and grouped
- Count word types and tokens
- Each frame defines a single category

Frequent frames

Evaluation

$$\text{Accuracy} = \frac{\text{hits}}{\text{hits} + \text{false alarms}}$$

All pairs of tokens compared in each frame-based category

$\Rightarrow$ measures proportion of all words grouped together that were grouped correctly

Frequent frames

Evaluation

$$\text{Accuracy} = \frac{\text{hits}}{\text{hits} + \text{false alarms}}$$

All pairs of tokens compared in each frame-based category

⇒ measures proportion of all words grouped together that were grouped correctly

$$\text{Completeness} = \frac{\text{hits}}{\text{hits} + \text{misses}}$$

All pairs compared across all categorized tokens

⇒ measures degree to which frames group tokens that belong to same word class

Frequent frames

Evaluation

$$\text{Accuracy} = \frac{\text{hits}}{\text{hits} + \text{false alarms}}$$

All pairs of tokens compared in each frame-based category

$\Rightarrow$ measures proportion of all words grouped together that were grouped correctly

$$\text{Completeness} = \frac{\text{hits}}{\text{hits} + \text{misses}}$$

All pairs compared across all categorized tokens

$\Rightarrow$ measures degree to which frames group tokens that belong to same word class

Coverage: percentage of tokens in corpus categorized by frames

Frequent frames

Results

Child	Accuracy		Completeness		Coverage	Categorized
	Frame	Rand	Frame	Rand		
Peter	0.98	0.49	0.06	0.03	48%	6%
Eve	0.98	0.51	0.06	0.03	46%	5%
Nina	0.98	0.48	0.08	0.04	51%	8%
Naomi	0.97	0.48	0.07	0.03	38%	5%
Anne	0.98	0.37	0.08	0.03	54%	4%
Aran	0.97	0.44	0.08	0.04	61%	5%
Mean	0.98	0.46	0.07	0.03	50%	6%

Adapted from Mintz 2003

Frequent frames

Some reservations

Frequent frames

Some reservations

1. High accuracy due to many single-type categories (e.g., want __ put $\longrightarrow$ {to})
   - Accuracy stable for high type-variability

Frequent frames

Some reservations

1. High accuracy due to many single-type categories (e.g., want __ put $\longrightarrow$ {to})
   - Accuracy stable for high type-variability
2. Absolute number of frequent frames per corpus
   - Similar results for relative frame frequencies

Frequent frames

Some reservations

1. High accuracy due to many single-type categories (e.g., want __ put $\longrightarrow$ {to})
   - Accuracy stable for high type-variability
2. Absolute number of frequent frames per corpus
   - Similar results for relative frame frequencies
3. Different frame-based categories might belong to bigger class
   - Unification with threshold for lexical overlap (e.g, $\theta = 20\% \rightsquigarrow$ 0.90 accuracy, 0.93 completeness)

Frequent frames

Conclusions

- Frequent frames induce extremely robust categories
- Low completeness due to frame-based categorization
- High coverage from categorizing small percentage of tokens
- Simple and psycholinguistically plausible computations
- Superior to previous models (e.g., Cartwright & Brent '97, Redington, Chater & Finch '98)

Frequent frames

Points of criticism

Frequent frames

Points of criticism

- Frequent-frame categories evaluated against tagged corpora
  - Tagging might not reflect categories children use
  - Tag-sets theory-dependent

Frequent frames

Points of criticism

- Frequent-frame categories evaluated against tagged corpora
    - Tagging might not reflect categories children use
    - Tag-sets theory-dependent
- Not clear how frequent-frame categories integrated into language processor model
    - encapsulated system for categorization only
    - frame-based categories carry no syntactic information

Frequent frames

Points of criticism

- Frequent-frame categories evaluated against tagged corpora
  - Tagging might not reflect categories children use
  - Tag-sets theory-dependent
- Not clear how frequent-frame categories integrated into language processor model
  - encapsulated system for categorization only
  - frame-based categories carry no syntactic information
- Approach has not been validated cross-linguistically
  - e.g., Erkelenz (UvA) shows that frame-based categories align with Dutch categories only 40%–71%

General perspective

Integration

Difficult to integrate statistical learning, psycholinguistic research and tools of computational linguistics:

Computational linguistics	Psycholinguistics
Learning from tagged corpora	Untagged input
Language specific algorithms	Typological viability
Large corpora (WSJ, Brown)	Child-directed speech (CHILDES)
Gold standard evaluation	Developmental data
Strong theoretical assumptions	Explanatory generality

BIG task (Chang, Lieven, Tomasello 2008)

Basic idea

- Incrementally generate sentences from unordered bag of words
- Learner predicts one word at a time using syntactic knowledge
- Recursive task, target word removed from bag of words

BIG task (Chang, Lieven, Tomasello 2008)

Basic idea

- Incrementally generate sentences from unordered bag of words
- Learner predicts one word at a time using syntactic knowledge
- Recursive task, target word removed from bag of words

Evaluation

- Sentence prediction success: target utterance predicted exactly
- Accuracy: percentage success over all utterances in test corpus

Statistical learners

BIG-SPA task suitable to compare statistical learners of syntax:

$C(w_{n-k} \ldots w_n)$	Frequency of n-gram $w_{n-k} \ldots w_n$ in input $(k = 0, 1, 2)$
NW	Number of word tokens in corpus
$\mathrm{Ch}(w_n)$	Choice function for word w_n
Bigram	$\mathrm{Ch}(w_n) = C(w_{n-1}, w_n)/C(w_{n-1})$
Trigram	$\mathrm{Ch}(w_n) = C(w_{n-2}, w_{n-1}, w_n)/C(w_{n-2}, w_{n-1})$
Bigram + Trigram	$\ldots$
Unigram + BG + TG	$\mathrm{Ch}(w_n) = C(w_n)/\mathrm{NW} + \ldots$
Backed-off TG	TG if > 0, else BG if > 0, else UG

BIG-SPA procedure

BIG-SPA procedure

1. Split input corpus into training/test set (90%/10%).

BIG-SPA procedure

1. Split input corpus into training/test set (90%/10%).
2. Collect learner statistics from training set.

BIG-SPA procedure

1. Split input corpus into training/test set (90%/10%).
2. Collect learner statistics from training set.
3. For each utterance u in test set, create bag of words b.

BIG-SPA procedure

1. Split input corpus into training/test set (90%/10%).
2. Collect learner statistics from training set.
3. For each utterance u in test set, create bag of words b.
4. For each word nw in u: for each word w in b, calculate Choice(w).

BIG-SPA procedure

1. Split input corpus into training/test set (90%/10%).
2. Collect learner statistics from training set.
3. For each utterance u in test set, create bag of words b.
4. For each word nw in u: for each word w in b, calculate Choice(w).
5. Add w with highest Choice(w) to *newu*.

BIG-SPA procedure

1. Split input corpus into training/test set (90%/10%).
2. Collect learner statistics from training set.
3. For each utterance *u* in test set, create bag of words *b*.
4. For each word *nw* in *u*: for each word *w* in *b*, calculate Choice(w).
5. Add *w* with highest Choice(w) to *newu*.
6. Remove *nw* from *b*, repeat until $b = \emptyset$.

BIG-SPA procedure

1. Split input corpus into training/test set (90%/10%).
2. Collect learner statistics from training set.
3. For each utterance u in test set, create bag of words b.
4. For each word nw in u: for each word w in b, calculate Choice(w).
5. Add w with highest Choice(w) to $newu$.
6. Remove nw from b, repeat until $b = \emptyset$.
7. If $newu = u$, increment SPA count by 1.

Typologically-different corpora

12 corpora from CHILDES:
Cantonese, Croatian, English, Estonian, French, German, Hebrew, Hungarian, Japanese, Sesotho, Tamil, Welsh

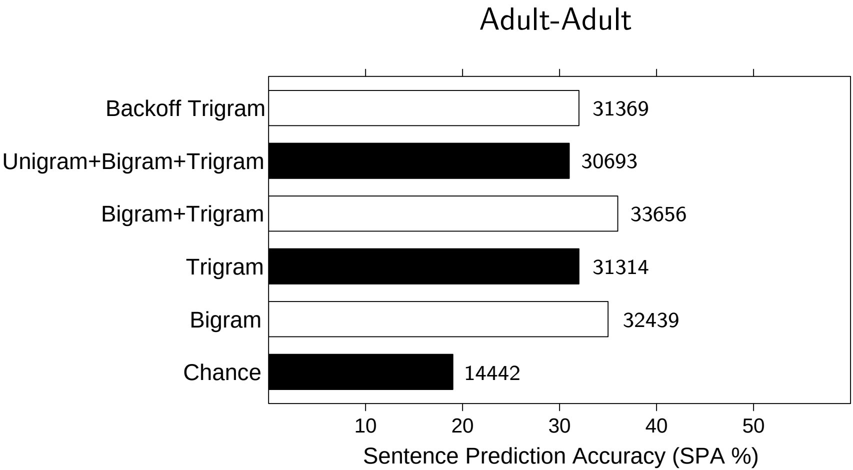
Four common word orders:
SVO (English), SOV (Japanese), VSO (Welsh), No dominant order (Hungarian)

Rigid (less rigid) word order:
English, French, Cantonese (German, Japanese, Croatian, Hungarian, Tamil)

Argument omission: Japanese, Cantonese

Rich morphology: Croatian, Estonian, Hungarian

BIG-SPA results



Adapted from Chang, Lieven & Tomasello, 2008

A psycholinguistically motivated learner

Adjacency-prominence learner

$C(w_{n-1}, w_n)$	Frequency of bigram $w_{n-1}w_n$
$P(w_a, w_b)$	Frequency that word w_a occurred before w_b in an utterance at any distance
$\text{Pair}(w_a, w_b)$	Frequency that words w_a, w_b occurred together in same sentence in any order
Length	Number of words in bag-of-word

Adjacency	$\text{Ch}_{adj}(w_n) = C(w_{n-1}, w_n)/\text{Pair}(w_{n-1}, w_n)$
Prominence	$\text{Ch}_{pro}(w_n) = \sum_{w_b} P(w_n, w_b)/\text{Pair}(w_n, w_b)$ where w_b are all the words in bag (except w_n)
Adjacency-Prominence	$\text{Ch}(w_n) = \text{Length} \times \text{Ch}_{adj}(w_n) + \text{Ch}_{pro}(w_n)$

Comparison

Sentence production constrained by syntactic & semantic factors

- Syntactic constraints: Adjacency statistics (normalized bigram)
- Semantic constraints: Prominence statistics (more prominent message components tend to be produced earlier)

Adjacency-prominence learner achieves significantly higher score than any other learner:

SPA 46% (48994 utterances across corpora, Adult-Adult)

Conclusions

Conclusions

- BIG-SPA task does not require gold standard for syntax

Conclusions

- BIG-SPA task does not require gold standard for syntax
- Can be used for typologically-different languages

Conclusions

- BIG-SPA task does not require gold standard for syntax
- Can be used for typologically-different languages
- Allows comparison of learning algorithms in theory-neutral way

Conclusions

- BIG-SPA task does not require gold standard for syntax
- Can be used for typologically-different languages
- Allows comparison of learning algorithms in theory-neutral way
- Allows to detect typological biases of particular algorithms

Conclusions

- BIG-SPA task does not require gold standard for syntax
- Can be used for typologically-different languages
- Allows comparison of learning algorithms in theory-neutral way
- Allows to detect typological biases of particular algorithms
- Helps to integrate psycholinguistic modelling and methods from computational linguistics

References

Chang, F., Lieven, E., and Tomasello, M. (2008). Automatic evaluation of syntactic learners in typologically-different languages. *Cognitive Systems Research*, *9*, 198–213.

Gerken, L., Wilson, R., and Lewis, W. (2005). 17-month-olds can use distributional cues to form syntactic categories. *Journal of Child Language*, *32*, 249–268.

Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*, 431–436.

Mintz, T. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*, 91–117.

Redington, M., Chater, N., and Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22(4)*, 425–469.